DOCUMENT RESUME

ED 353 282 TM 019 000

AUTHOR Marsh, Herbert W.
TITLE A Longitudinal Perspective of Students' Evaluations of University Teaching: Ratings of the Same Teachers over a 13-Year Period.
PUB DATE Apr 92
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Covariance; Cohort Analysis; College Students; Factor Analysis; Factor Structure; Foreign Countries; Generalization; Higher Education; Longitudinal Studies; *Professors; *Profiles; Student Attitudes; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; *Test Reliability
IDENTIFIERS Australia; *Rater Reliability; *Students Evaluations of Educational Quality

ABSTRACT

Students' evaluations of teacher effectiveness are explored as a multidimensional construct, emphasizing the Students' Evaluations of Educational Quality (SEEQ) instrument developed by H. W. Marsh (1987). An overview is presented of studies in which Marsh evaluated longitudinal data from an archive of responses to nearly 1 million SEEQ instruments representing 50,000 courses collected over 13 years. The focus is on a cohort of nearly 200 teachers evaluated over the period. The analysis of studies considers: (1) the generalizability of the SEEQ factor structure; (2) higher order factor structure; (3) the generalizability of ratings over time; (4) models of covariance stability; and (5) analyses of teacher profiles. Analyses provided clear support for the generalizability of the SEEQ factor structure over time, over courses, and over teaching at different levels. Higher order factor analyses suggest that higher order factors that might underlie SEEQ factors are not particularly useful in providing a smaller number of scores with which to summarize SEEQ responses. Each instructor was found to have a reasonably distinct profile. Mean ratings of the same teachers evaluated consistently showed no systematic increases or decreases. An appendix contains the SEEQ instrument. Four tables and three figures present study data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
HERBERT W. MARSH
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED353282

A LONGITUDINAL PERSPECTIVE OF STUDENTS' EVALUATIONS OF UNIVERSITY TEACHING: RATINGS OF THE SAME TEACHERS OVER A 13-YEAR PERIOD

Herbert W. Marsh, University of Western Sydney, Macarchur

Paper presented at the 1992 Annual Meeting of the American Educational Research Association, San Francisco, April 20-24

TM019000

BEST COPY AVAILABLE

A Longitudinal Perspective of Students' Evaluations of University Teaching: Ratings of the Same Teachers Over a 13-Year Period

Herbert W. Marsh, University of Western Sydney, Macarthur

Paper presented at the 1992 Annual Meeting of the American Educational Research Association, San Francisco, April 20-24.

INTRODUCTION

This paper examines students' evaluations of teaching effectiveness (SETs) as a multidimensional construct and emphasizes the Students' Evaluations of Educational Quality (SEEQ; Marsh, 1987) instrument. Previous research demonstrates that class-average SETs are: 1) multidimensional; 2) reliable and stable; 3) primarily a function of the instructor who teaches a course rather than the course that is taught; 4) relatively valid against a variety of indicators of effective teaching; 5) relatively unaffected by a variety of variables hypothesized as potential biases; and 6) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions.

Despite the huge volume of SET research, most research has considered ratings collected in one specific course on a single occasion and there is surprisingly little longitudinal research that considers the ratings of the same teachers over an extended period of time. For example, inferences about how SETs are related to teacher age are typically based on cross-sectional studies. Cross-sectional studies, however, provide a poor basis for inferring what ratings younger teachers will receive later in their careers or what ratings older teachers would have received if evaluated earlier in their careers. Clearly, there are important limitations in the use of cross-sectional data for evaluating how ratings of the same instructor varies over time.

In the present investigation I present an overview of studies in which I evaluated longitudinal data derived from an archive of responses to nearly 1 million SEEQ instruments representing 50,000 courses that have been collected over a 13-year period of time. I begin by briefly reviewing support for the generalizability of the SEEQ factor structure. For purposes of the longitudinal analyses I focus on a cohort of nearly 200 teachers who were evaluated continuously over the 13-year period. In different studies I examine three perspectives of the question of how ratings of the same instructors vary over time.

1. How well does the SEEQ factor structure generalize across teaching at different levels and in different disciplines (Marsh & Hocevar, 1991a; ; also see Marsh & Roche, in press)? This question is addressed by the comparison of 22 factor analyses of ratings of unique groups of teachers that vary in terms of academic discipline (e.g., psychology, Spanish, engineering) and level (e.g., undergraduate and graduate level courses). In related research I have also examined second-order factors (Marsh, 1991b, 1991c).

2. How do the mean ratings of the same set of teachers vary over time (Marsh & Hocevar, 1991b)? The question here is whether the mean ratings for the longitudinal cohort of teachers systematically increase or decrease over time. Contrary to the results from reviews of cross-sectional studies, this analysis shows that the mean ratings of these teachers are stable over time.

3. How highly related are ratings from different occasions over the 13-year period (Marsh, 1991)? Here I looked at the test-retest correlations from one year to the next and for longer periods of time. The results showed that test-retest correlations were high for short periods of time and were nearly as high for much longer periods of time. Different theoretical models positing a "simplex" pattern of growth and single latent construct with no systematic change were evaluated.

4. How stable is the profile of SEEQ factors for the same teacher over extended periods of time (Marsh & Bailey, 1991)? For any particular set of ratings, ratings for one scale (e.g., Enthusiasm) will be higher or lower than another scale (e.g., Organization). Here I consider whether the profile of SEEQ scores for the same teacher generalizes over courses, over course

levels (graduated and undergraduate classes), and over the 13-year period during which the ratings were collected. These results show that the profile of SEEQ scores is also very stable over time, more stable apparently than even the overall ratings.

THE GENERALIZABILITY OF THE SEEQ FACTOR STRUCTURE

The selection and revision of the SEEQ items was based on literature reviews, student and teacher responses about the importance of items, teacher responses about the usefulness of items, examination of open-ended comments by students, and psychometric properties of the responses, thus supporting the content validity of SEEQ responses. Factor analytic support for the SEEQ scales is particularly strong. To date, more than 30 published factor analyses of SEEQ responses have identified the 9 factors that SEEQ is designed to measure.

Insert Tables 1 and 2 About Here

Marsh and Hocevar (1991a) described the archive of SEEQ responses that contains ratings of 50,000 classes (representing responses to nearly 1 million SEEQ surveys). From this archive, 24,158 courses were selected and classified into one of 21 different subgroups (see Table 1) varying in terms of teacher rank (teaching assistant or regular staff), level of instruction (undergraduate or graduate), and academic discipline. Twenty-two separate factor analyses of the total sample (see Table 2) and each of the subsamples all identified the 9 factors that SEEQ is designed to measure, providing very strong support for the generality of the factor structure underlying SETs.

For each course, two sets of factor scores were derived: one based on the factor analysis of the total sample of 24,158 courses and one based on the specific subsample (of the 21 subsamples) to which the course was classified. These two sets of factor scores were correlated in each of the 21 different subgroups. High correlations among factor scores representing the same factor provide support for the comparability of the different factor structures. Nearly all of the 189 correlations (9 SEEQ factor x 21 subsamples) were greater than .95 and the majority were larger than .99.

Because of the psychometric properties of the SEEQ instrument and because of the size and diversity of the data base considered here, the results provide stronger support for the generality of the factor structure underlying SETs than does any previous research.

HIGHER-ORDER FACTOR STRUCTURES

SEEQ clearly measures distinct dimensions of teaching effectiveness, but some researchers argue that SETs can be explained by one or a relatively few number of higher-order factors that incorporate distinct first-order factors. I (Marsh, 1991a) examined this possibility in a higher-order factor analyses based on an application of confirmatory factor analysis (HCFA). Based on previous research, different models were posited that had 9 first-order factors and either 1, 2, 3, or 4 higher-order factors. The model with 9 first-order factors and 4 higher-order factors fit the data the best. Even this model was not entirely satisfactory in that much of the true score variance in the first-order factors could not be explained in terms of the higher-order factors. The study demonstrates that the SEEQ responses cannot be explained adequately by one or even a few summary scores and illustrates the application of hierarchical confirmatory factor analysis.

Insert Figure 1 About Here

GENERALIZABILITY OF RATINGS OVER TIME

The two most common approaches to the study of stability and change refer to the stability of means over time (mean stability) and to the stability of individual differences over time (covariance stability or test-retest correlations). In this section I focus primarily on mean stability.

In an early review of research based largely on primary and secondary teaching, Ryans (1960) reported an overall negative relation between teaching experience and teaching

effectiveness. He suggested an initial increase in effectiveness during the first few years, a leveling out period, and then a period of gradual decline. In her review of research since the early 1960s, Barnes (1985) reached a similar conclusion. At the university level, Feldman (1983) reviewed studies relating overall and content-specific dimensions of SETs to teacher age, teaching experience, and academic rank. He reported that SETs were only weakly related to these three measures of seniority, but that distinct patterns were evident. Overall evaluations tended to be negatively correlated with age and -- to a lesser extent -- years of teaching experience, but tended to be positively correlated with academic rank. Thus, younger teachers, teachers with less teaching experience, and teachers with higher academic ranks tended to receive somewhat higher evaluations. Age and teaching experience showed reasonably similar patterns of correlations with overall and content-specific dimensions. Academic rank, however, tended to be positively correlated with some characteristics such as subject knowledge, intellectual expansiveness, and value of course materials, but negatively correlated with other characteristics such as class discussion, respect for students, helpfulness and availability to students. Consistent with the reviews by Ryans (1960) and Barnes (1985), Feldman noted that in the few studies that specifically examined nonlinear relations, there was some suggestion of an inverted U-shaped relation in which ratings improved initially, peaked at some early point, and then declined slowly thereafter.

As I noted earlier, there are important limitations in the use of cross-sectional data for evaluating how ratings of the same instructor varies over time. For this reason, I examined changes in ratings of a large number of teachers who had been evaluated continuously over a 13-year period with SEEQ. Using the SEEQ archive I selected all teachers who were evaluated at least once during each of 10 different years over the 13 year period the ratings were collected. This process identified 195 different teachers who had been evaluated in a total of 6024 different courses (an average of 30.9 classes per teacher) from a total of 31 different academic departments. A multiple regression approach to ANOVA was used in which linear and nonlinear effects of year (1976-1988), course level (2 = graduate, 1 = undergraduate), and their interaction were evaluated. Whereas graduate level courses tended to be rated higher than undergraduate level courses, the effect of year -- the major focus of this analysis -- was consistently small. First-order correlations and standardized beta weights reflecting changes in ratings over time varied from -.067 to +.016 for the 9 SEEQ factors and the two overall rating items.

The most important influence in the SETs is the instructor. In order to evaluate the influence of the instructor the mean rating of each instructor over all undergraduate classes and over all graduate classes was computed. In the main regression models considered, these instructor mean ratings were included along with the linear and nonlinear components of the year (1976-1988), the course level, and their interactions. Hence, the effects of the individual instructor were controlled in evaluating the effects of the other variables.

Insert Table 3 About Here

The individual instructor accounted for most of the variance in each of the different SEEQ scores (Table 3). Because the instructor effect reflects ratings of the same teacher over time, this demonstrates the covariance stability (test-retest correlations) of the SETs. There were almost no systematic changes in the mean rating across all teachers over time, thus supporting the mean stability of the ratings. Year accounted for no more than 1/4 of 1% in any of the evaluation scores, and -- despite the large N and powerful design -- only reached statistical significance for 2 of 11 scores. Supplemental analyses suggested that the standards that students used apparently did not change over this period. The nonlinear effects suggested from cross-sectional studies were not observed for either the total sample, or subsamples of teachers with little, intermediate, or substantial amounts of teaching experience at the start of the 13-year longitudinal study. These results are important because this is apparently the only study to examine the mean stability of faculty ratings using a longitudinal design with a large and diverse group of teachers over such a long period of time. The results also demonstrate the potential dangers in trying to infer changes of over time from cross-sectional data instead of true longitudinal data.

MODELS OF COVARIANCE STABILITY

The purpose of analyses described in this section is to evaluate models of covariance stability and change. Data are SEEQ responses from a cohort of 157 teachers who had been

evaluated in at least one course during each of at least 7 years during an 8-year period. Separate analyses were conducted on the 7 SEEQ scales measured by 4 items (the other 2 scales are measured by 2 or 3 items).

Insert Figure 2

The simplex model is apparently the most widely used approach for analyzing covariance stability in longitudinal data. The critical feature of the simplex model is that correlations between measures collected at adjacent occasions are highest and that sizes of correlations decrease steadily as a function of the number of occasions separating two measures. A viable alternative is a simple "one factor" model in which ratings of the same SEEQ scale at different points in time reflects a stable "true" score and an error component. The simplex model implies systematic gradual changes over time such that the longer the intervening period of time the lower the test-retest correlations whereas the one factor model does not necessarily imply any systematic change. Inspection of correlations between ratings on different occasions, however, suggests that the simplex model may be wrong. Thus, for example, ratings of Learning/Value in 1980 and 1981 correlated .58 whereas correlations between ratings in 1980 and 1987 were .56.

Simplex and one-factor models were both evaluated with single-indicator models (the mean of the 4 items used to infer each scale) and multiple-indicator models (the four items). In both cases, the one-factor models fit the data better than the simplex model. The results challenge a potential over-reliance on the simplex model. The results also suggest that the assumption of a systemic, gradual change in SETs of the same teacher over time is apparently false.

GENERALIZABILITY: PROFILE ANALYSIS

Thus far, I have considered only the generalizability of individual SEEQ scales. I noted (Marsh, 1987), the need to examine profiles of SEEQ scores as well as the individual scales that make up the profile. More specifically I suggested that each instructor has a distinguishable profile of SEEQ scales (e.g., high on organization and low on enthusiasm) that generalizes over different course offerings and is distinct from the profiles of other instructors. Because apparently no other research known has evaluated SET profiles in this manner, I conducted a profile analysis of ratings selected from the longitudinal SEEQ archive. In this study I considered 3079 sets of class-average responses for 123 instructors -- an average of 25 classes per instructor -- who had been evaluated regularly over a 13-year period for both graduate and undergraduate courses. Because there were so many sets of ratings for each instructor, it was possible to determine a characteristic profile of SEEQ scores for each instructor. In profile analyses, it is important to distinguish between the level of scores (whether an instructor consistently receives high or low ratings) and the shape of the profile (e.g., relatively higher on organization and relatively lower on enthusiasm).

The profiles of four teachers (figure 3) illustrate the "level" and "shape" comparisons that are the focus of the profile analysis. Each profile is the average rating across all sets of ratings of the same teacher collected during the 13-year period. Instructors 1 and 2 have generally higher ratings than instructors 3 and 4, demonstrating the effect of level. The effect of shape can be seen by comparing the Enthusiasm and Organization scores for the different teachers. Instructors 1 and 3 have consistently higher ratings for Organization than Enthusiasm, whereas teachers 2 and 4 have consistently higher ratings for Enthusiasm than for Organization.

Insert Figure 3

In the ANOVA model used to evaluate profiles, I considered the main effects of instructors, course level, instructor by course-level interaction, and the extent to which these varied depending on the particular SEEQ component. These main effects reflect the effects averaged across the 9 SEEQ scales.

The very large effect of the instructor is an important finding. The eta squared value of .371 is equivalent to an average correlation of .61 between ratings of the same instructor (averaged across all SEEQ factors) on different course offerings. This is consistent with my earlier (Marsh, 1981) finding that the average correlation was .71 across two offerings of the same course and .52 across ratings of different courses.

Although not a main focus the present investigation, the results again show that graduate courses are evaluated somewhat more highly than undergraduate courses. The statistically significant, although modest instructor by course-level interaction suggests that some teachers get systematically higher ratings in graduate level courses whereas other teachers get higher

ratings in undergraduate courses. This interaction effect, however, is much smaller than the main effect of instructor.

The most important effect in these analyses is the extent to which the effect of instructor varies as a function of the particular SEEQ factor. If this interaction is small, it suggests that highly rated teachers receive consistently high ratings across all SEEQ factors and that poorly rated teachers receive consistently poor ratings across all SEEQ factors. In fact, however, the effect of the instructor varies systematically with the SEEQ factor. Of the total variation in profiles in 3,000 classes, nearly 50% is due to the specific instructor. This substantial effect clearly demonstrates that the profiles associated with each instructor are consistent across different course offerings by that instructor, and distinct from the profiles of other instructors. This finding is the most important result of the Marsh and Bailey (1991) study.

The profile of 9 SEEQ scales (e.g., Enthusiasm, Organization, Group Interaction) for each instructor was shown to be distinct from the profiles of other instructors, generalized across course offerings over the 13-year period, and generalized across undergraduate and graduate level courses. This support for the existence of a distinct profile that is specific to each instructor has important implications for the use of SETs and opens up new, largely unexplored areas for further research. For example, validity research typically focuses on the relations between external validity to either overall SETs, or, sometimes, to specific SET scales. Because there are reliable individual differences in SET profiles, these results imply that researchers should also consider how different profiles are related to external criteria of effective teaching. For example, it may be that student learning is maximized when both Enthusiasm and Organization are higher, whereas being high on just one or the other is insufficient. Similarly, the demonstration of systematic, reliable individual differences in SET profiles supports the use of profiles in formative feedback and summative evaluations, and has important implications for the study of teaching and teaching styles. The results also provide further support for the multidimensionality of SETs.

SUMMARY AND IMPLICATIONS

This set of studies has important theoretical and practical implications for the use of students evaluations. The unique strength of the set of studies is the large number of classes evaluated with the same SEEQ instrument over such a long period of time. The series of factor analyses provided clear support for the generalizability of SEEQ factor structure over time, over courses in different disciplines, and over teaching at different levels. Higher-order factor analyses suggested that whereas there may be higher-order factors underlying the 9 SEEQ factors, these factors were apparently not particularly useful in providing a smaller number of scores with which to summarize SEEQ responses. Consistent with these findings, the profile analysis indicated that each instructor has a reasonably distinct profile of SEEQ scores that generalizes over time and across courses taught at both graduate and undergraduate levels.

In two studies I looked at the stability of mean differences and the covariance stability of individual differences. In contrast to suggestions based on reviews of cross-sectional studies, the mean ratings of the same cohort of teachers who were evaluated consistently over a 13-year period showed no systematic increases or decreases. This study also indicated that a substantial portion of the variance in the ratings could be explained in terms of the teacher who taught the course. In the study of covariance stability, simplex models that posit systematic changes in the ratings over time were not able to fit the data very well, whereas a good fit was found with a simple one-factor model positing that ratings for each year reflect a single stable factor that generalizes over time.

References

Barnes, J. (1985). Experience and student achievement/teacher effectiveness. In T. Husen & T. N. Postlethwaite (Eds.), International encyclopedia of education: Research and studies (pp. 5125-5128). Oxford: Pergamon Press.

Feldman, K. A. (1983). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. Research in Higher Education, 18, 3-124.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-388 (Whole Issue No. 3).

Marsh, H. W. (1991a). Covariance stability in multiwave Panel Studies: Comparison of Simplex Models and One-factor Models (in review).

Marsh, H. W. (1991b). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia (1991). Journal of Educational Psychology, 83, 416-421.

Marsh, H. W. (1991c). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. Journal of Educational Psychology, 83, 285-296.

Marsh, H. W., & Bailey, M. (in press). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. Journal of Higher Education.

Marsh, H. W., & Dunkin, M. J. (in press). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (ed.), Higher Education: Handbook of Theory and Research (vol. 8). New York: Agathon.

Marsh, H. W., & Hocevar, D. (1991a). Multidimensional students' evaluations of teaching effectiveness: Factor structure stability across academic discipline, instructor level, and course level. Teaching and Teacher Education, 7, 9-18.

Marsh, H. W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. Teaching and Teacher Education 7, 303-314.

Marsh, H. W., & Roche, L. A. (in press). The use of students' evaluations of university teaching in different settings: The applicability paradigm. Australian Journal of Education.

Ryans, D. G. (1960). Prediction of teacher effectiveness. In C. W. Harris (Ed.), Encyclopedia of educational research (pp. 1486-1491). New York: Macmillan.

Appendix

The Students' Evaluations of Educational Quality (SEEQ) Instrument. Interested individuals are invited to use the SEEQ freely for their own personal use. Parties interested in using SEEQ on a wider basis are requested to obtain permission from Herbert W. Marsh, which will be freely given. This will allow the application and performance of SEEQ to be monitored. Some "Student and Course Characteristics," labelling, and, perhaps, aspects of the instructions may be changed as necessary, but in other respects the SEEQ instrument should be left intact.

© 1976 HERBERT MARSH Printed in U.S.A. NCS Trans-Optic® MP08-80189-32 A2302

USC EVALUATION SERVICES

USE NO. 2 PENCIL ONLY

• COMPLETELY ERASE ANY RESPONSE YOU WISH TO CHANGE

SELECT THE BEST RESPONSE FOR EACH OF THE FOLLOWING STATEMENTS, LEAVING A BLANK ONLY IF IT IS CLEARLY NOT RELEVANT

		VERY POOR		MODERATE (AVERAGE)		VERY GOOD
1	**LEARNING:** YOU FOUND THE COURSE INTELLECTUALLY CHALLENGING AND STIMULATING	1	2	3	4	5
2	YOU HAVE LEARNED SOMETHING WHICH YOU CONSIDER VALUABLE	1	2	3	4	5
3	YOUR INTEREST IN THE SUBJECT HAS INCREASED AS A CONSEQUENCE OF THIS COURSE	1	2	3	4	5
4	YOU HAVE LEARNED AND UNDERSTOOD THE SUBJECT MATERIALS IN THIS COURSE	1	2	3	4	5
5	**ENTHUSIASM:** INSTRUCTOR WAS ENTHUSIASTIC ABOUT TEACHING THE COURSE	1	2	3	4	5
6	INSTRUCTOR WAS DYNAMIC AND ENERGETIC IN CONDUCTING THE COURSE	1	2	3	4	5
7	INSTRUCTOR ENHANCED PRESENTATIONS WITH THE USE OF HUMOR	1	2	3	4	5
8	INSTRUCTOR'S STYLE OF PRESENTATION HELD YOUR INTEREST DURING CLASS	1	2	3	4	5
9	**ORGANIZATION:** INSTRUCTOR'S EXPLANATIONS WERE CLEAR	1	2	3	4	5
10	COURSE MATERIALS WERE WELL-PREPARED AND CAREFULLY EXPLAINED	1	2	3	4	5
11	PROPOSED OBJECTIVES AGREED WITH THOSE ACTUALLY TAUGHT SO YOU KNEW WHERE COURSE WAS GOING	1	2	3	4	5
12	INSTRUCTOR GAVE LECTURES THAT FACILITATED TAKING NOTES	1	2	3	4	5
13	**GROUP INTERACTION:** STUDENTS WERE ENCOURAGED TO PARTICIPATE IN CLASS DISCUSSIONS	1	2	3	4	5
14	STUDENTS WERE INVITED TO SHARE THEIR IDEAS AND KNOWLEDGE	1	2	3	4	5
15	STUDENTS WERE ENCOURAGED TO ASK QUESTIONS & WERE GIVEN MEANINGFUL ANSWERS	1	2	3	4	5
16	STUDENTS WERE ENCOURAGED TO EXPRESS THEIR OWN IDEAS AND/OR QUESTION THE INSTRUCTOR	1	2	3	4	5
17	**INDIVIDUAL RAPPORT:** INSTRUCTOR WAS FRIENDLY TOWARD INDIVIDUAL STUDENTS	1	2	3	4	5
18	INSTRUCTOR MADE STUDENTS FEEL WELCOME IN SEEKING HELP/ADVICE IN OR OUTSIDE OF CLASS	1	2	3	4	5
19	INSTRUCTOR HAD A GENUINE INTEREST IN INDIVIDUAL STUDENTS	1	2	3	4	5
20	INSTRUCTOR WAS ADEQUATELY ACCESSIBLE TO STUDENTS DURING OFFICE HOURS OR AFTER CLASS	1	2	3	4	5
21	**BREADTH:** INSTRUCTOR CONTRASTED THE IMPLICATIONS OF VARIOUS THEORIES	1	2	3	4	5
22	INSTRUCTOR PRESENTED THE BACKGROUND OR ORIGIN OF IDEAS/CONCEPTS DEVELOPED IN CLASS	1	2	3	4	5
23	INSTRUCTOR PRESENTED POINTS OF VIEW OTHER THAN HIS/HER OWN WHEN APPROPRIATE	1	2	3	4	5
24	INSTRUCTOR ADEQUATELY DISCUSSED CURRENT DEVELOPMENTS IN THE FIELD	1	2	3	4	5
25	**EXAMINATIONS:** FEEDBACK ON EXAMINATIONS/GRADED MATERIALS WAS VALUABLE	1	2	3	4	5
26	METHODS OF EVALUATING STUDENT WORK WERE FAIR AND APPROPRIATE	1	2	3	4	5
27	EXAMINATIONS/GRADED MATERIALS TESTED COURSE CONTENT AS EMPHASIZED BY THE INSTRUCTOR	1	2	3	4	5
28	**ASSIGNMENTS:** REQUIRED READINGS/TEXTS WERE VALUABLE	1	2	3	4	5
29	READINGS, HOMEWORK, ETC. CONTRIBUTED TO APPRECIATION AND UNDERSTANDING OF SUBJECT	1	2	3	4	5
30	**OVERALL:** COMPARED WITH OTHER COURSES YOU HAVE TAKEN AT USC, THIS COURSE WAS . . .?	1	2	3	4	5
31	**OVERALL:** COMPARED WITH OTHER INSTRUCTORS YOU HAVE HAD AT USC, THIS INSTRUCTOR WAS . . .?	1	2	3	4	5
	STUDENT AND COURSE CHARACTERISTICS: (LEAVE BLANK IF NO RESPONSE APPLIES)					
32	COURSE DIFFICULTY, RELATIVE TO OTHER COURSES, WAS (1-VERY EASY . . . 3-MEDIUM . . . 5-VERY HARD)	1	2	3	4	5
33	COURSE WORKLOAD, RELATIVE TO OTHER COURSES, WAS (1-VERY LIGHT . . 3-MEDIUM . . . 5-VERY HEAVY)	1	2	3	4	5
34	COURSE PACE WAS (1-TOO SLOW . . . 3-ABOUT RIGHT . . . 5-TOO FAST)	1	2	3	4	5
35	HOURS PER WEEK REQUIRED OUTSIDE OF CLASS 1) 0 - 2, 2) 2 - 5, 3) 5 - 7, 4) 7 - 12, 5) OVER 12	1	2	3	4	5
36	LEVEL OF INTEREST IN THE SUBJECT PRIOR TO THIS COURSE (1-VERY LOW . . . 3-MEDIUM . . . 5-VERY HIGH)	1	2	3	4	5
37	OVERALL GPA AT USC 1) BELOW 2.5, 2) 2.5 TO 3.0, 3) 3.0 TO 3.4, 4) 3.4 TO 3.7, 5) ABOVE 3.7 LEAVE BLANK IF NOT YET ESTABLISHED AT USC	1	2	3	4	5
38	REASON FOR TAKING THE COURSE (1-MAJOR REQUIRE., 2-MAJOR ELECTIVE, 3-GENERAL ED REQUIRE., 4-MINOR/RELATED FIELD, 5-GENERAL INTEREST ONLY)—SELECT THE ONE WHICH IS BEST	1	2	3	4	5
39	YEAR IN SCHOOL 1) FRESH. 2) SOPH. 3) JR. 4) SR. 5) GRAD	1	2	3	4	5
40	EXPECTED GRADE IN THE COURSE (1-A, 2-A-, 3-B+, 4-B, 5-B-, 6-C+, 7-C, 8-C-, 9-D, 10-F)	1	2	3	4	5
		6	7	8	9	10
41	MAJOR DEPARTMENT 1) SOC SCI/COMM. 2) NAT SCI/MATH. 3) HUMANITIES. 4) BUSINESS. 5) EDUCATION	1	2	3	4	5
	6) ENGINEERING. 7) PERF. ARTS. 8) PUB. AFFAIRS. 9) OTHER. 10) UNDECLARED/UNDECIDED	6	7	8	9	10

SUPPLEMENTAL QUESTIONS (USE RESPONSES BELOW FOR INSTRUCTOR'S QUESTIONS)

42 ①②③④⑤	46 ①②③④⑤	50 ①②③④⑤	54 ①②③④⑤
43 ①②③④⑤	47 ①②③④⑤	51 ①②③④⑤	55 ①②③④⑤
44 ①②③④⑤	48 ①②③④⑤	52 ①②③④⑤	56 ①②③④⑤
45 ①②③④⑤	49 ①②③④⑤	53 ①②③④⑤	57 ①②③④⑤

INSTRUCTOR'S NAME DEPARTMENT NAME COURSE NUMBER

INSTRUCTOR'S NAME	DEPARTMENT NAME	COURSE NUMBER

INSTRUCTIONS

This evaluation form is intended to measure your reactions to this instructor and course. Results will be reported to USC Administrators to be used as part of the overall evaluation of the instructor. These evaluations will have salary and promotion implications so please take them very seriously. When you have finished, a designated student will pick up the evaluations and take them to a drop off point for processing. Your responses will remain anonymous and summaries will not be given to the instructor until after grades have been assigned.

****Put Instructor's Name, Department Name and Course Number at top
(i.e., Smith, Psychology, 200)

OPEN-ENDED COMMENTS

PLEASE INDICATE THE IMPORTANT CHARACTERISTICS OF THIS INSTRUCTOR/COURSE WHICH HAVE BEEN MOST VALUABLE TO YOUR LEARNING EXPERIENCE.

1 _____

_____

_____

2 _____

_____

_____

3 _____

_____

PLEASE INDICATE CHARACTERISTICS OF THIS INSTRUCTOR/COURSE WHICH YOU FEEL ARE MOST IMPORTANT FOR HIM/HER TO WORK ON IMPROVING (PARTICULARLY ASPECTS NOT COVERED BY RATING ITEMS)

1 _____

_____

_____

2 _____

_____

_____

3 _____

_____

_____

PLEASE USE THE ADDITIONAL SPACE TO CLARIFY ANY OF YOUR RESPONSES OR TO MAKE OTHER COMMENTS.

_____

_____

Table 1 (Reproduced From Marsh & Hocevar, 1991a)

Summary of the 21 Subsamples of Courses

No. of classes	Academic unit
Undergraduate courses taught by teaching assistants	
1. 431	General
2. 610	Business
3. 565	Humanities
4. 1606	Social Sciences
5. 683	Spanish and Portuguese
6. 1368	Economics
7. 902	Communication
Undergraduate courses taught by regular faculty	
1. 1421	General
2. 2326	Business
3. 956	Humanities
4. 2320	Social Sciences
5. 1693	Engineering
6. 590	History
7. 538	Psychology
Graduate courses taught by regular faculty	
1. 757	General
2. 2049	Business
3. 1157	Social Sciences
4. 957	Engineering
5. 1213	Education
6. 457	Systems Engineering
7. 1559	Safety and Systems Management
Total	
24,158	

Note. For present purposes all classes with six or more sets of ratings were classified into 21 subsamples such that each subsample had at least 400 classes. All classes were first categorized into general groups consisting of classes taught by teaching assistants, undergraduate classes taught by regular faculty, and graduate courses taught by regular faculty. Classes were then classified into divisions or schools (e.g., Social Sciences or Engineering) and then into specific departments (e.g., Psychology or Systems Engineering) whenever there were more than 400 classes. All classes were classified into one and only one subsample.

Table 2 (Reproduced From Marsh & Hocevar, 1991a)

Factor Analysis Results for the Total Sample of 24,158 Sets of Class-average Responses: Factor Loadings and Factor Correlations

SEEQ scales and items (paraphrased)	SEEQ factors								
	Lrn	Enth	Orgn	Grp	Ind	Brd	Exam	Asgn	Work
Learning/Value									
Course challenging & stimulating	*.434*	.168	.103	.015	.014	.159	.099	.155	.291
Learned something valuable	*.607*	.083	.100	.026	.050	.103	.085	.147	.113
Increase subject interest	*.646*	.078	.034	.039	.058	.169	.074	.131	.020
Learned & understood subject matter	*.487*	.043	.176	.152	.045	.047	.112	.149	−.217
Overall course rating	*.410*	.211	.173	.041	.042	.085	.166	.175	.069
Instructor Enthusiasm									
Enthusiastic about teaching	.095	*.544*	.129	.072	.195	.115	.052	.069	.025
Dynamic and energetic	.064	*.714*	.094	.059	.085	.083	.069	.071	.042
Enhanced presentation with humor	.089	*.650*	−.023	.103	.078	.129	.090	.054	−.045
Teaching style held your interest	.137	*.581*	.187	.131	.026	.050	.110	.073	.017
Overall instructor rating	.172	*.392*	.245	.083	.141	.096	.140	.075	.039
Organization/Clarity									
Lecturer explanations clear	.146	.165	*.510*	.176	.060	.075	.079	.104	−.072
Materials well explained & prepared	.069	.087	*.677*	.060	.075	.073	.094	.118	.005
Course objectives stated & pursued	.128	.026	*.529*	.055	.070	.065	.175	.184	.024
Lectures facilitated taking notes	.031	.040	*.589*	−.093	.049	.175	.146	.044	.020
Group Interaction									
Encouraged class discussion	.058	.103	.011	*.769*	.070	.033	.067	.080	.002
Students shared knowledge/ideas	.066	.049	−.015	*.797*	.095	.093	.048	.073	−.029
Encouraged questions & gave answers	.059	.105	.167	*.583*	.151	.094	.100	.080	.001
Encouraged expression of ideas	.045	.069	.035	*.674*	.182	.110	.094	.070	−.013
Individual Rapport									
Friendly towards individual students	.051	.163	−.001	.176	*.612*	.063	.112	.057	−.038
Welcomed students seeking help/advice	.042	.059	.061	.078	*.786*	.036	.093	.059	−.007
Interested in individual students	.086	.140	.001	.137	*.647*	.057	.138	.059	.004
Accessible to individual students	−.014	−.028	.139	.037	*.636*	.099	.136	.104	.010
Breadth of coverage									
Contrasted various implications	.043	.037	.118	.059	.068	*.676*	.077	.109	.065
Gave background of ideas/concepts	.087	.085	.134	.020	.044	*.662*	.056	.122	.004
Gave different points of view	.035	.066	.086	.123	.101	*.636*	.097	.113	−.004
Discussed current developments	.207	.113	.018	.086	.039	*.562*	.084	.040	.000
Examinations/Grading									
Examination feedback valuable	.034	.039	.111	.047	.101	.028	*.670*	.088	.044
Evaluation methods fair/appropriate	.047	.044	.011	.043	.107	.078	*.749*	.099	−.033
Tested course content as emphasized	.063	.036	.129	.034	.064	.047	*.643*	.146	−.026
Assignments/Readings									
Readings/texts were valuable	−.008	−.004	.019	.022	.018	.053	.025	*.885*	−.003
They contributed to understanding	.127	.021	.036	.027	.039	.012	.140	*.716*	.072
Workload/Difficulty									
Course difficulty (easy-hard)	−.028	.030	.051	−.059	−.017	.096	.015	.018	*.861*
Course workload (light-heavy)	.100	−.054	.004	.085	−.001	.002	−.035	.038	*.907*
Course pace (slow-fast)	−.098	.101	.055	−.099	.005	−.001	.035	.040	*.689*
Hours per week outside of class	.148	−.044	−.085	.034	−.001	−.006	−.006	.042	*.798*
Factor pattern correlations									
Learning/Value	1.000								
Instructor Enthusiasm	.434	1.000							
Organization/Clarity	.407	.427	1.000						
Group Interaction	.350	.364	.210	1.000					
Individual Rapport	.263	.400	.331	.455	1.000				
Breadth of Coverage	.449	.419	.454	.327	.352	1.000			
Examinations/Grading	.401	.392	.511	.315	.493	.403	1.000		
Assignments/Readings	.488	.319	.431	.312	.338	.418	.510	1.000	
Workload/Difficulty	.128	.076	.044	−.072	−.009	.106	.033	.154	1.000

Note. Target loadings, the factor loadings items designed to define each SEEQ factor, are presented in italics.

Table 3 (Reproduced From Marsh & Hocevar, 1991b)

Changes in Multiple Dimensions of Students Evaluations Over Time for Ratings of the Same instructor: The Effects of Instructor, Year (1976-1988), Level (undergraduate and graduate), and their Interaction (N=3135).

		Standardized Beta Weights For:						
	r for							
Dimension	Instr	Instr	Year	Year2	Level	YrxLev	YR2xlev	Mult R
Factor Scores								
Learning/Value	.701**	.703**	.001	-.045**	-.023	.018	.025	.703**
Enthusiasm	.822**	.822**	-.016	-.019	-.003	.010	.006	.822**
Organization	.770**	.770**	-.048**	-.025	.000	.017	-.004	.772**
Group Interact	.814**	.815**	-.012	-.020	-.009	-.013	.010	.815**
Indiv Rapport	.747**	.746**	-.026	.016	.006	.006	-.009	.748**
Breadth	.735**	.735**	.005	-.011	.000	.009	-.007	.736**
Exams	.678**	.678**	-.028	-.017	.006	-.008	-.014	.678**
Assignments	.704**	.704**	-.004	-.024	-.008	.012	.006	.704**
Workload	.797**	.797**	-.020	-.009	.010	-.009	.007	.797**
Overall Ratings								
Course	.725**	.725**	-.031	-.028	-.013	-.031	.019	.726**
Instructor	.756**	.755**	-.048**	-.020	-.010	.009	.015	.758**

Note. The Instructor (instr) component was obtained by taking the mean of the instructor ratings for undergraduate classes and for graduate classes, and then including these means in the prediction of ratings. Because these means were computed separately for graduate and undergraduate level courses, it has the effect of eliminating variance due to course level.

* $p < .05$; ** $p < .01$.

Table 4 (Reproduced From Marsh & Bailey, 1991)

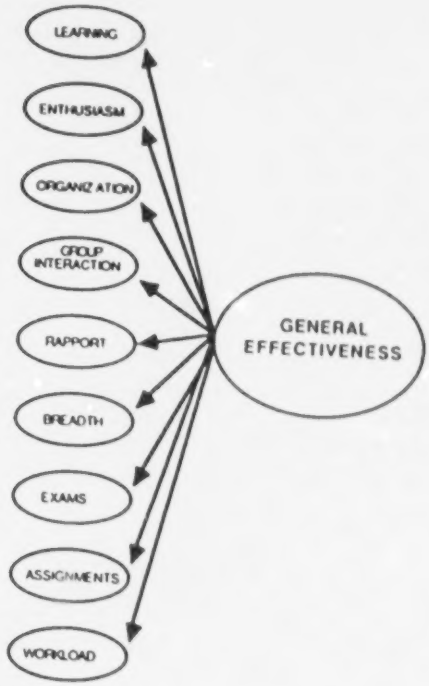
Univariate Repeated Measures and MANOVA Analyses of SEEQ Profiles

Repeated Measures Analysis					MANOVA			
			Effect Sizes[a]					
Source	SS	df	η^2	ω^2	Wilks' Lambda	Hypoth df	Error df	Mutivariate F
Between (Total Scores)								
Instructor (I)	3894	122	.371	.349				
Course Level (L)	166	1	.016	.016				
I x L	934	122	.089	.067				
Error Between	5493	2833						
Total Between	10488	3078						
Within (Profiles)								
Scales x I	8099	976	.470	.451	.0050	976	22594	23.60
Scales x L	108	8	.006	.006	.8923	8	2826	42.65
Scales x I x L	1195	976	.069	.050	.3403	976	22594	3.31
Error Within	7812	22664						
Total Within	17214	24632						

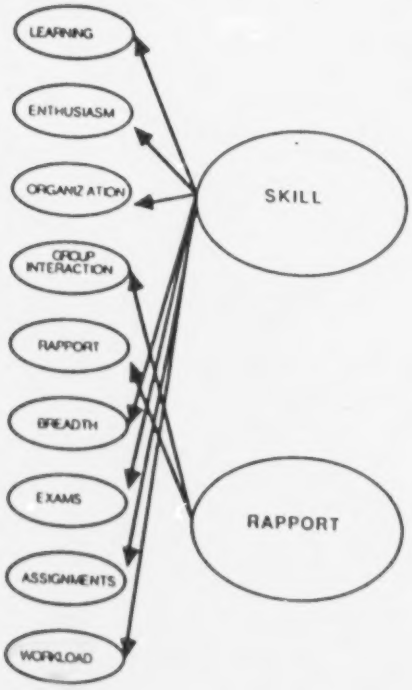
Note. The repeated measures and MANOVA approaches for the Total Scores (i.e., the Between Groups portion) are equivalent since there is only one dependent measure. For both approaches the 9 SEEQ scores were transformed into 8 difference scores between adjacent SEEQ factors -- the standard "repeated" transformation in SPSS (1988). Whereas the ordering of the SEEQ scores is arbitrary, the results in no way depend on the particular ordering used (see Tabachnick & Fidell, 1989, for further discussion).

[a] The Greenhouse-Geisser, Huynh-Feldt, and Lower-bound Epsilons were 0.872, 0.950, and 0.125 respectively. Even when the Lower-bound Epsilon that is known to be maximally conservative was used , all tests of statistical significance were significant at $p < .001$.

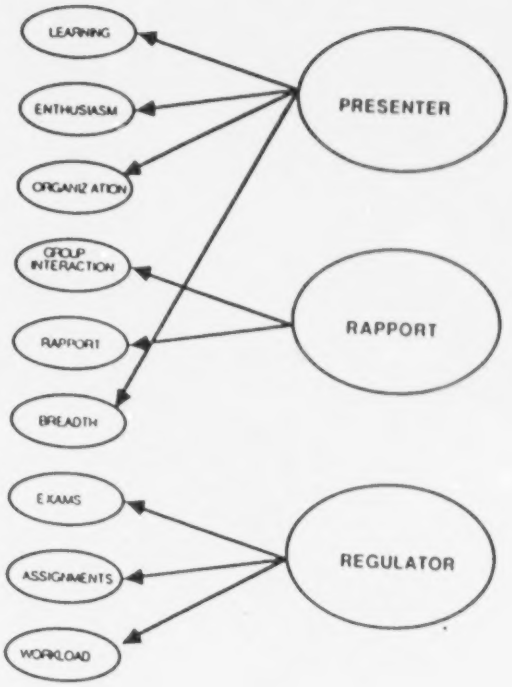
Figure 1 (Reproduced From Marsh, 1991b)



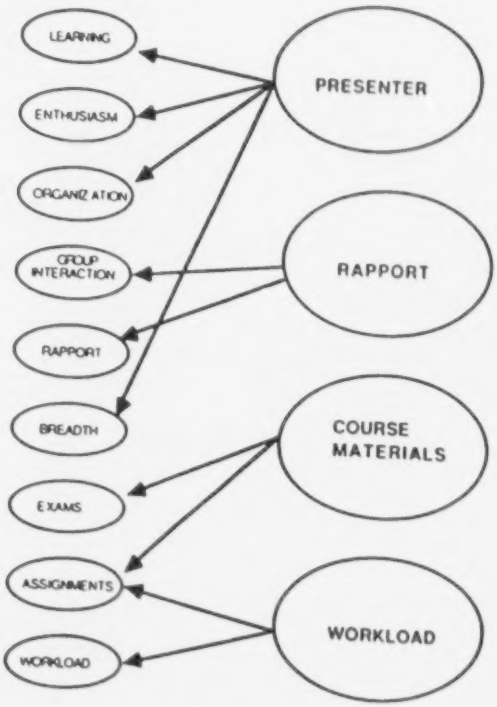
MODEL H1



MODEL H2



MODEL H3

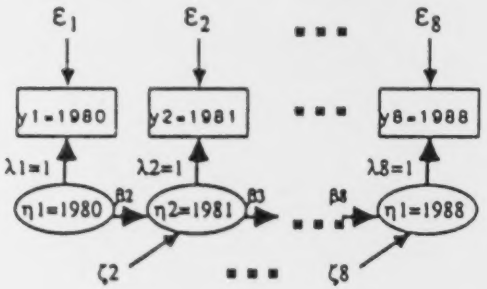


MODEL H4

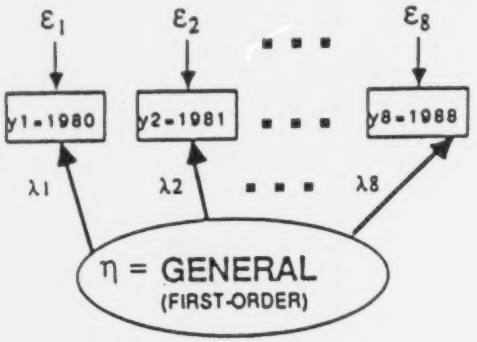
Figure 1 (opposite) Four a priori higher-order models of relations among first-order Students' Evaluations of Education Quality (SEEQ) factors (Each first-order factor is inferred from multiple indicators based on the design of the SEEQ. To avoid clutter, the multiple indicators of each first-order factor and correlations among higher-order factors are not presented.)

Figure 2 (Reproduced From Marsh, 1991a)

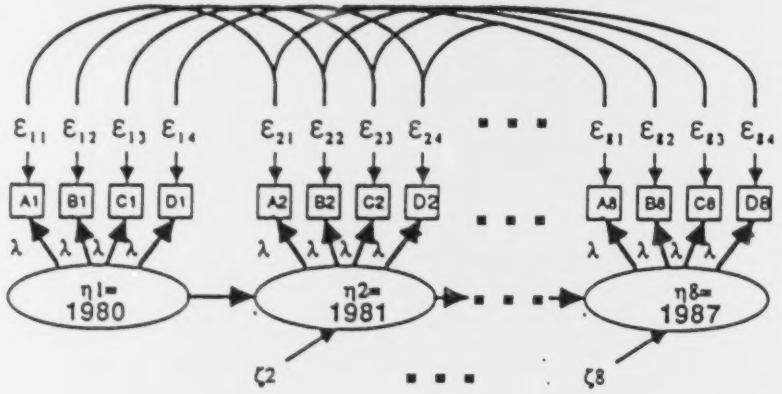
M2: QUASI-SIMPLEX (SINGLE INDICATORS)



M4: ONE FACTOR (SINGLE INDICATORS)



M3: SIMPLEX (MULTIPLE INDICATORS)



M6: FIRST-ORDER MUTIPLE INDICATOR MODEL

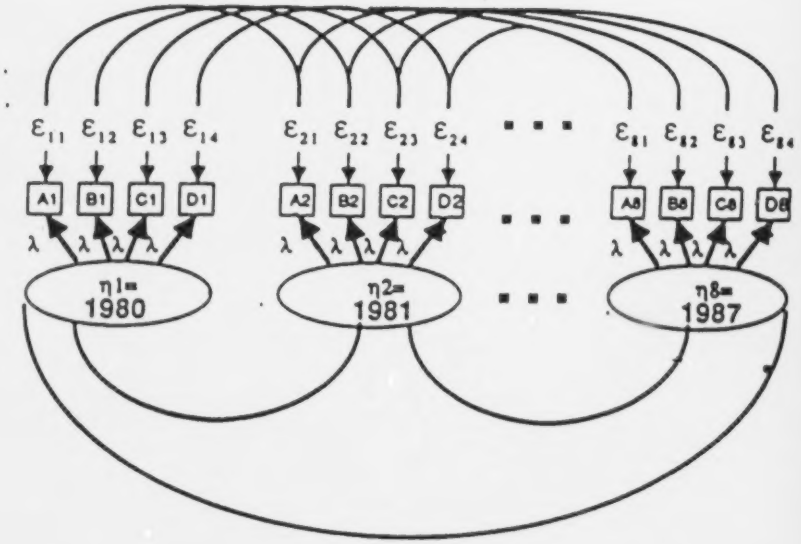


Figure 3 (Reproduced From Marsh & Bailey, 1991)

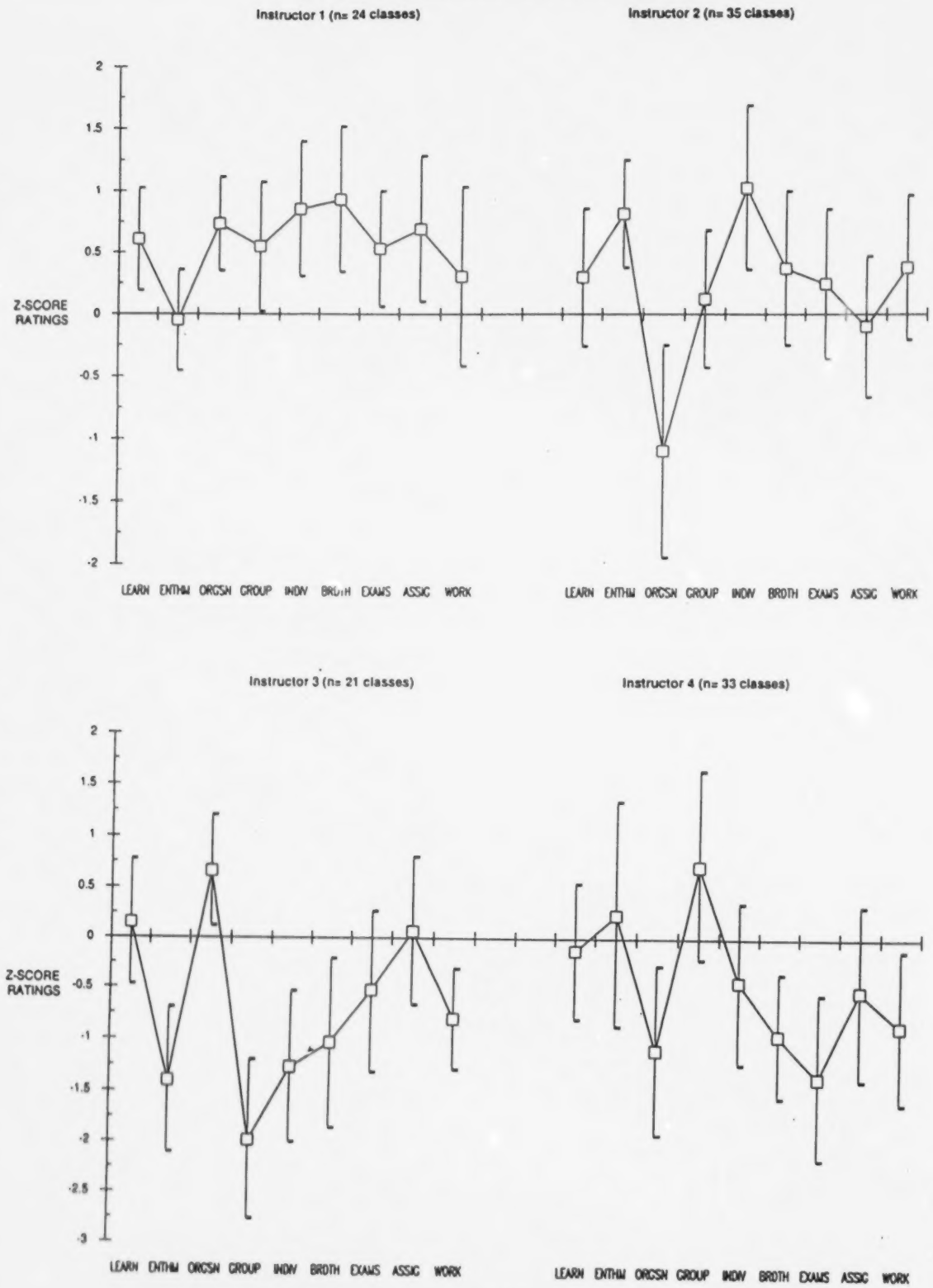


Figure 1. Profiles of nine SEEQ scales for 4 instructors (see Appendix for more detail on the SEEQ scales). All scores were standardized (mean = 0, SD = 1) across all sets of ratings used in the study. Each profile represents the mean score for each SEEQ scale (the boxes), averaged across all the classes for that instructor. Thus, for example, all scores above the line representing a z-score of zero reflect ratings that are above average. Also presented for each scale, is the range of scores corresponding to the mean plus and minus one standard deviation (based on the set of ratings for the particular instructor for that particular scale).